Extracting collocations in Russian: Statistics vs. Dictionary

Maria Khokhlova[1,2]

[1]Saint-Petersburg State University – Universitetskaya nab., 11
199034 Saint-Petersburg – Russia

[2]Institute for Linguistic Studies – Tuchkov per., 9 – 199053 Saint-Petersburg – Russia

Abstract

The notion of collocation is quite ambiguous. A concise survey of different approaches to it (British contextualism, lexicographical approach, approach of the "Meaning-Text" theory) is proposed in the paper. The paper discusses the results of retrieving collocations from a corpus of Russian texts. The data obtained is compared to the data given for set expressions in modern Russian dictionaries. The paper also explores the role of statistical measures for extracting collocations in Russian, and the issue of their applicability to the Russian language.

Keywords: collocation extraction, Russian, association measures, MI-score, t-score, log-likelihood ratio, dictionaries.

1. Introduction

Probabilistic nature of language is beyond any doubt. Thus statistical data is an important factor when describing different linguistic phenomena.

The methods for collocation extraction proposed in most works have not been evaluated so far whether they can be applicable to Russian, and if yes, to what degree. Also there's a question what types of set phrases they allow to retrieve. The explanatory dictionaries do not always consistently reflect the information about set phrases. The boundary between free and set phrases is quite ambiguous.

According to some scientists (Mel'chuk, 1960) the property of stability is inherent to all word combinations. A threshold of stability should be chosen to range them, above which a word combination can be called a set phrase.

The term "collocation" has come to use in Russian linguistics, after Western linguistics, to designate set phrases. Although the term itself appeared long ago (Akhmanova, 1966), it is not generally recognized by Russian scholars. Such language units have various names in different works; cf. "set verbal-noun expressions" (Deribas, 1983), "analytic lexical collocations" (Teliya, 1996) etc. The majority of authors understand under collocation a statistically set phrase. Collocations can be put between free phrases and idioms on a scale of phrases.

At first the notion of collocation was introduced by the founder of London School of Structural Linguistics and the representative of British contextualism J.R. Firth (Firth, 1957). The word meaning, in Firth's opinion, is closely connected with its ability to collocability.

Collocation is a tendency of a word to a certain environment. So, he stated the hypothesis according to which it is possible for a word to be attributed to a group by its neighbourhood. The parts of collocation occupy certain positions and, thus, are characterized by mutual expectancy of appearance. Collocations can be viewed as forms of meaning (Firth, 1957).

It is possible to allocate also the lexicographic approach to studying the phenomenon of collocation. While in British contextualism collocation is defined on the basis of statistical assumptions about the probability of co-occurrence of two (or more) lexemes, and especially frequent combinations of lexical units are considered as collocations, the lexicographic approach considers collocation as a semantic-syntactic unit or a combination of lexically defined elements of grammatical structures. Within the framework of this approach a special attention is given to the structures that underlie collocations.

2. The notion of "collocation" in Russian linguistics

The monograph (Borisova, 1995a) has proved to be the first work in Russian linguistics, completely devoted to the research of the concept of collocation on a material of Russian. One of the key properties of collocation is "the impossibility of prediction of such combinations on the basis of meanings of their components" (Borisova, 1995a: 13).

Another classification of collocations is given in (Teliya, 1996). Under the term "collocation" Teliya understands a combination characterized by a nominative regularity, i.e. due to the bound component it has the ability to designate the senses possessing the content of common category, "typical of aspectual and temporal meanings and also of meanings correlating with semantic cases of deep structure (in the sense of Fillmore (Fillmore, 1968))" (Teliya, 1996). In Teliya's opinion, it is this principle that underlies lexical functions of the "Meaning-Text" theory. For example, *byt' ne v nastrojenii* = "to be in bad mood" (cf. *byt' v dome* = "to be in the house"), *luch nadezhdy* = "a ray of hope" (cf. *luch sveta* = "a ray of the sun"), *kormilo vlasti* = "at the helm" (cf. *kormilo korablya* = "helm of a ship") etc.

In the "Meaning-Text" theory collocations are considered as a subclass of more extensive class of set phrases, or phrasemes. "An idiom is an expression consisting of several lexemes whose meaning cannot be completely deduced by general rules of the given language from the meanings of its constituent lexemes, from the morphological characteristics (if those are available) assigned to them semantically and from their syntactic configuration by the general rules of the given language" (Iordanskaja, Mel'chuk, 2007: 215).

According to Mel'chuk and Teliya, collocations can be understood as word-combinations in which one of the elements is viewed as a semantic dominant, and another is chosen depending on it in order to express the sense of the whole combination (M. Hausmann, A. Cowie[1], S. Kahane and A. Polguère adhere to the same approach). The dependent word, thus, can be interpreted only in combination with the dominant. The similar standpoint we find in (Borisova, 1995a).

3. The analysis of retrieving collocations in Russian

Nowadays there are several ways in linguistics to calculate the degree of collocates' coherence. They are based on the comparison of frequencies for word pairs obtained on a material of a real corpus with independent (relative) frequencies. Statistically significant

[1] A. Cowie calls such combinations *restricted collocations*.

deviations of real frequencies from hypothetical probabilities (for more details see (Stubbs, 1995)) are searched.

Statistical methods for data on corpus structure treatment are widely used in corpus linguistics. There are different measures based on calculation of a degree of nearness of words in a text, namely, MI (mutual information), t-score, log-likelihood (henceforth LL), z-score, chi-square.

The object of research in the given work is collocations of Russian, and their presentation in dictionaries of modern Russian.

The aim was to carry out a number of experiments in order to find a suitable association measure for different classes of set phrases; to define opportunities of statistical methods as a whole and several measures in particular; to find ways of combination of statistical and semantic-syntactical methods in retrieving collocation.

We have led a series of experiments with the purpose of comparing the efficiency of statistical methods.

During experiment the following ideas were tested:

- to what degree the proposed methods can be applicable to Russian;
- whether the given methods allow to reveal other classes of set phrases.

We have chosen the collocations of 19 nouns that don't have homonyms as material for our research. The nouns have been selected on the principle of their sufficient high frequency (see the Electronic Frequency Dictionary of Russian by S. Sharoff (Sharoff, 2002)): *власть* "power", *внимание* "attention", *возможность* "opportunity", *война* "war", *вопрос* "question", *дождь* "rain", *жизнь* "life", *закон* "law", *любовь* "love", *место* "place", *мнение* "opinion", *мысль* "thought", *ночь* "night", *ответ* "answer", *помощь* "help", *радость* "joy", *слово* "word", *случай* "case", *смысл* "sense".

The research has been led on the corpus of Russian newspapers created at the University of Leeds (Great Britain)[2] under the guidance of S. Sharoff. This corpus includes around 78 million words from several major Russian newspapers (for example, "Izvestia"), its part-of-speech tagging was done using the program Mystem[3].

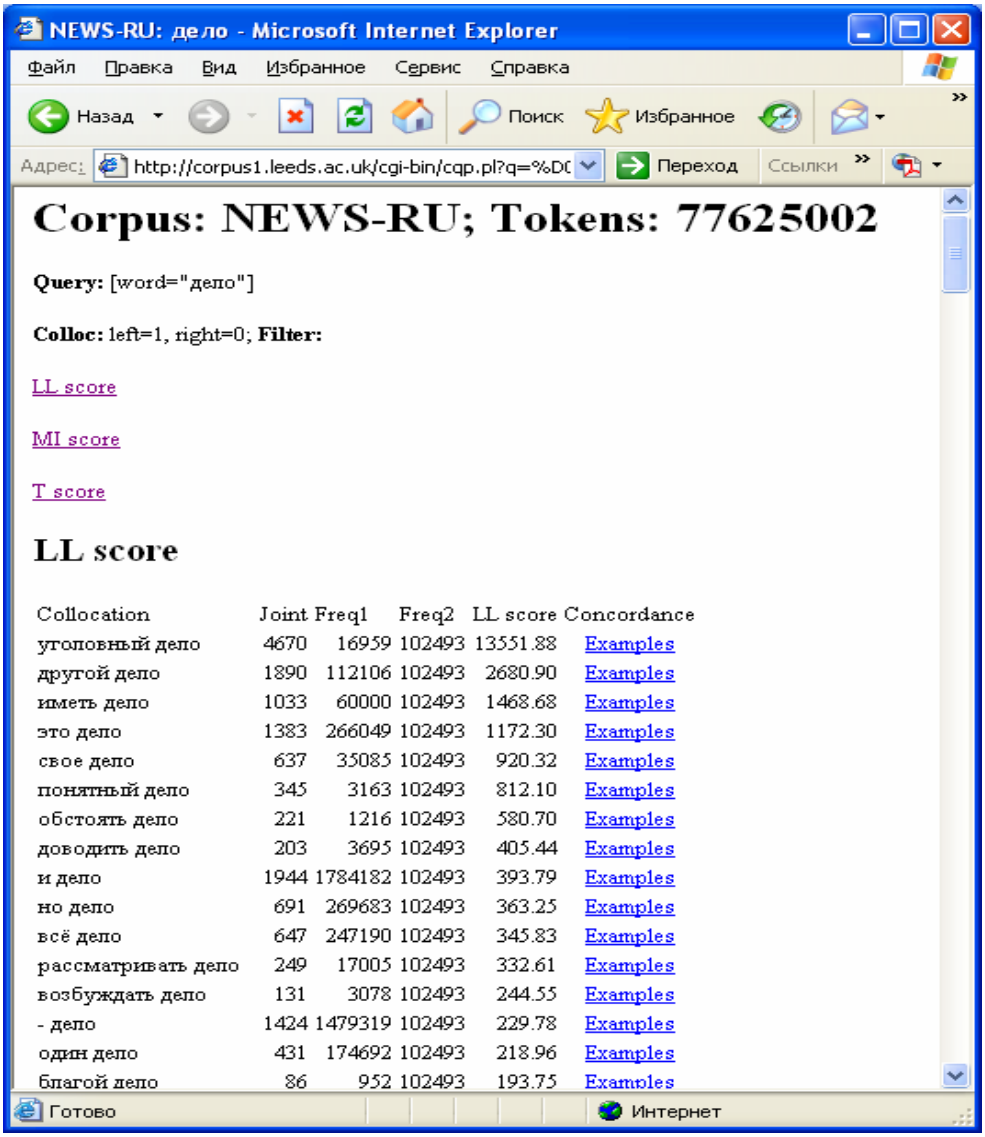
In a search mode one can choose one or several statistical measures (MI, t-score, LL), set a span in words, and also it is also possible to set a part of speech of a collocate.

It is necessary to mention two moments beforehand. First, each element of the corpus which stands before or after a blank including punctuation marks is considered a token. Secondly, the corpus manager CQP uses lemmas while processing data, thus, results of a search are presented by combinations of lemmas.

The result of the query is represented by a list of collocations organized in the form of one, two or three tables (depending on the quantity of the chosen measures) with six data columns (see Figure 1):

[2] http://corpus.leeds.ac.uk/ruscorpora.html

[3] http://corpora.narod.ru/mystem



Corpus: NEWS-RU; Tokens: 77625002

Query: [word="дело"]

Colloc: left=1, right=0; **Filter:**

LL score

MI score

T score

LL score

Collocation	Joint	Freq1	Freq2	LL score	Concordance
уголовный дело	4670	16959	102493	13551.88	Examples
другой дело	1890	112106	102493	2680.90	Examples
иметь дело	1033	60000	102493	1468.68	Examples
это дело	1383	266049	102493	1172.30	Examples
свое дело	637	35085	102493	920.32	Examples
понятный дело	345	3163	102493	812.10	Examples
обстоять дело	221	1216	102493	580.70	Examples
доводить дело	203	3695	102493	405.44	Examples
и дело	1944	1784182	102493	393.79	Examples
но дело	691	269683	102493	363.25	Examples
всё дело	647	247190	102493	345.83	Examples
рассматривать дело	249	17005	102493	332.61	Examples
возбуждать дело	131	3078	102493	244.55	Examples
- дело	1424	1479319	102493	229.78	Examples
один дело	431	174692	102493	218.96	Examples
благой дело	86	952	102493	193.75	Examples

Figure 1. Example of the output of the query on the word дело "business"

The first column shows the collocation (represented by lemmas) itself. The joint frequency of occurrence of bigram's components, the frequency of the first word and the frequency of the second word stand in the second, third and fourth columns accordingly.

The data in all tables were sorted on decrease of value of a corresponding measure. The query results for each noun were brought to one table. We compared them to the entries for these nouns in the Dictionary of Collocations (Borisova, 1995b), in the explanatory dictionaries of Russian (the Dictionary of Modern Russian (Slovar' sovremennogo russkogo literaturnogo jazyka, 1948-1965); the Big Academy Dictionary of Russian (Bol'shoj akademicheskij slovar' russkogo yazyka, 2004-2007), the Dictionary of Russian (Slovar' russkogo jazyka, 1957-1961)) and in the Dictionary of Synonyms and Similar Expressions (Abramov, 2006).

3.1. Results for Log-Likelihood

For LL measure the following results were received. 1763 bigrams were found in total. Among them there were:

47 bigrams are fixed in two or more dictionaries;

79 bigrams are fixed only in (Borisova, 1995b);

48 bigrams are fixed only in (Slovar' russkogo jazyka, 1957-1961);

20 bigrams are fixed only in (Abramov, 2006);

11 bigrams are fixed in (Bol'shoj akademicheskij slovar' russkogo jazyka, 2004-2007);

6 bigrams are fixed only in (Slovar' sovremennogo russkogo literaturnogo jazyka, 1948-1965).

Also there were 15 combinations with punctuation marks.

Values of LL proved to be the largest for the collocations found in two or more dictionaries.

№	Collocation	Joint	Freq1	Freq2	LL score	Concordance
1.	обращать внимание (pay attention)	4118	12455	19714	14361.30	Examples
2.	этот вопрос (this question)	4684	476434		5130.73	Examples
3.	на вопрос (to the question)	5887	1105092		4786.25	Examples
4.	давать возможность (enable)	1904	60300		3892.76	Examples
5.	особый внимание (special attention)	1427	16112	19714	3848.17	Examples
6.	иметь место (take place)	1899	60000		3568.69	Examples
7.	весь жизнь (the whole life)	2161	130350	59718	3441.61	Examples
8.	в ответ (in response)	3543	2534398		3419.57	Examples
9.	е место (corpus failure)	1307	9896		3411.37	Examples
10.	общественный мнение (public opinion)	1066	18429		2841.35	Examples
11.	иметь возможность (have a chance)	1439	60000		2731.97	Examples
12.	привлекать внимание (attract attention)	971	9401	19714	2687.61	Examples
13.	рассматривать вопрос (consider the question)	1242	17005		2572.59	Examples
14.	первый место (the first place)	1665	111613		2499.13	Examples
15.	оказывать помощь (help)	977	17711		2491.59	Examples
16.	решать вопрос (solve a question)	1486	47147		2446.03	Examples
17.	высказывать мнение (express an opinion)	774	8475		2239.46	Examples

№	Collocation	Joint	Freq1	Freq2	LL score	Concordance
18.	федеральный закон (federal law)	1026	37679	49277	2152.57	Examples
19.	второй место (the second place)	1208	45762		2150.41	Examples
20.	в ночь (at night)	2363	2534398		2098.42	Examples
21.	такой мнение (such an opinion)	1227	150108		2066.85	Examples
22.	всякий случай (any case)	711	11480		2062.06	Examples
23.	свое мнение (own opinion)	853	35085		1892.07	Examples
24.	третий место (the third place)	833	19293		1686.15	Examples
25.	на место (into place)	2506	1105092		1539.12	Examples
26.	медицинский помощь (medical aid)	574	10352		1459.52	Examples
27.	получать возможность (get an opportunity)	932	79406		1430.13	Examples
28.	новогодний ночь (New Year's night)	419	2690		1410.19	Examples
29.	принимать закон (pass a law)	833	68313	49277	1409.33	Examples
30.	задавать вопрос (ask a question)	568	4990		1305.35	Examples

Table 1. The first 30 significant collocations according to LL

3.2. Results for MI

1755 bigrams were found in total. Among them there were:

68 bigrams fixed in two or more dictionaries;

73 bigrams fixed only in (Borisova, 1995b);

27 bigrams are fixed only in (Slovar' russkogo jazyka, 1957-1961);

13 bigrams are fixed only in (Abramov, 2006);

9 bigrams are fixed in (Bol'shoj akademicheskij slovar' russkogo jazyka, 2004-2007);

25 bigrams are fixed only in (Slovar' sovremennogo russkogo literaturnogo jazyka, 1948-1965).

Also there were 11 combinations with punctuation marks.

№	Collocation	Joint	Freq1	Freq2	MI score	Concordance
1.	накрапывать дождь (drizzle)	7	19		14.95	Examples
2.	моросить дождь (drizzle)	19	67		14.57	Examples
3.	мифогенный любовь (mythogenic love)	4	4		14.39	Examples
4.	проливный дождь (downpour)	48	206		14.29	Examples
5.	метеорный дождь (meteor shower)	4	32		13.39	Examples
6.	варфоломеевский ночь (massacre of St. Bartholomew)	12	16		13.25	Examples
7.	метеоритный дождь (meteorite shower)	4	39		13.11	Examples
8.	вальпургиев ночь (Walpurgis-night)	5	8		12.99	Examples
9.	неослаблять внимание (give attention)	5	5	19714	12.57	Examples
10.	сакцентировать внимание (place emphasis)	5	5	19714	12.57	Examples
11.	утвердительный ответ (affirmative answer)	43	76		12.44	Examples
12.	здравый смысл (common sense)	240	1066		12.35	Examples
13.	замолвить слово (put in a word)	13	37		12.28	Examples
14.	нечаянный радость (unexpected joy)	20	219		12.18	Examples
15.	закрадываться мысль (creep, about a thought)	11	87		12.05	Examples
16.	кратковременный дождь (light rain)	32	702		11.94	Examples
17.	мелькнуть мысль (flit, about a thought)	17	146		11.93	Examples
18.	неразделять любовь (undivided love)	12	68		11.89	Examples
19.	крамольный мысль (rebellious thought)	12	106		11.89	Examples

№	Collocation	Joint	Freq1	Freq2	MI score	Concordance
20.	развернутый ответ (detailed answer)	11	30		11.82	Examples
21.	акцентировать внимание (place emphasis)	204	349	19714	11.79	Examples
22.	узурпировать власть (usurp power)	20	54		11.76	Examples
23.	шальной мысль (crazy thought)	12	118		11.74	Examples
24.	бытовать мнение (there is an opinion)	131	346		11.71	Examples
25.	платонический любовь (Platonic love)	6	39		11.69	Examples
26.	преступать закон (violate the law)	67	136	49277	11.59	Examples
27.	июльский дождь (July rain)	9	299		11.34	Examples
28.	лить дождь (pour, about rain)	21	753		11.23	Examples
29.	бессонный ночь (white night)	18	103		11.15	Examples
30.	однополый любовь (unisexual love)	13	128		11.09	Examples

Table 2. The first 30 significant collocations according to MI

Values of the MI measure are the largest for the collocations found only in (Slovar' russkogo jazyka, 1957-1961), and also found in two or more dictionaries. After examination of the list of results we found out, that only two combinations were retrieved (and both were not fixed in the dictionary of collocations) within a range from 0 to 1 (according to the value of MI). It allows us making a conclusion that the combination is statistically insignificant if the MI appears in the given interval. Thus the hypothesis that was applied to other languages can be extrapolated to Russian.

3.3. Results for t-score

1755 bigrams were found in total. Among them there were:

71 bigrams fixed in two or more dictionaries;

73 bigrams fixed only in (Borisova, 1995b);

22 bigrams are fixed only in (Slovar' russkogo iazyka, 1957-1961);

14 bigrams are fixed only in (Abramov, 2006);

8 bigrams are fixed in (Bol'shoi akademicheskii slovar' russkogo iazyka, 2004-2007);

23 bigrams are fixed only in (Slovar' sovremennogo russkogo literaturnogo iazyka, 1948-1965).

Also there were 20 combinations with punctuation marks.

№	Collocation	Joint	Freq1	Freq2	T score	Concordance
1.	на вопрос (to the question)	5887	1105092		70.20	Examples
2.	этот вопрос (this question)	4684	476434		65.28	Examples
3.	обращать внимание (pay attention)	4118	12455	19714	64.14	Examples
4.	в ответ (in response)	3543	2534398		55.19	Examples
5.	весь жизнь (the whole life)	2161	130350	59718	45.72	Examples
6.	в ночь (at night)	2363	2534398		44.60	Examples
7.	на место (into place)	2506	1105092		43.65	Examples
8.	давать возможность (enable)	1904	60300		43.34	Examples
9.	иметь место (take place)	1899	60000		43.18	Examples
10.	первый место (the first place)	1665	111613		40.01	Examples
11.	решать вопрос (solve a question)	1486	47147		37.99	Examples
12.	особый внимание (special attention)	1427	16112	19714	37.71	Examples
13.	иметь возможность (have a chance)	1439	60000		37.60	Examples
14.	на помощь (in help)	1613	1105092		36.48	Examples
15.	е место (corpus failure)	1307	9896		36.07	Examples
16.	рассматривать вопрос (consider the question)	1242	17005		35.02	Examples
17.	такой мнение (such an opinion)	1227	150108		34.54	Examples
18.	второй место (the second place)	1208	45762		34.37	Examples
19.	общественный мнение (public opinion)	1066	18429		32.59	Examples
20.	свой жизнь (own life)	1164	205621	59718	32.48	Examples
21.	федеральный закон (federal law)	1026	37679	49277	31.84	Examples
22.	оказывать помощь (help)	977	17711		31.18	Examples
23.	привлекать внимание (attract attention)	971	9401	19714	31.11	Examples

№	Collocation	Joint	Freq1	Freq2	T score	Concordance
24.	получать возможность (get an opportunity)	932	79406		29.98	Examples
25.	быть возможность (there is an opportunity)	1127	664975		29.39	Examples
26.	свое мнение (own opinion)	853	35085		29.07	Examples
27.	третий место (the third place)	833	19293		28.67	Examples
28.	принимать закон (pass a law)	833	68313	49277	28.48	Examples
29.	на жизнь (for a lifetime)	1342	1105092	59718	28.42	Examples
30.	во внимание (into account)	835	103853	19714	28.30	Examples

Table 3. The first 30 significant collocations according to t-score

The combinations that have large values of t-score prove to be rather frequent while, unlike the previous measures, one of their parts is a preposition or a pronoun. And also there were more bigrams (in comparison with other measures) in which a punctuation mark is one of their parts. Eg.: *война)* “war)”, *война ?*““war ?””, *война »*, “war »” etc.

We confirmed the hypothesis that t-score allows to retrieve collocations which have very frequent words, and also punctuation marks as their constituents. Thus, as well as for other languages, it is true for Russian that words with the largest value of t-score are frequent and can be combined with a large number of words. The right context reveals more combinations with punctuation marks than the left one.

3.4. Evaluation

The analysis of the data received shows that the majority of collocations (phrasemes), fixed in dictionaries, stand in the top part of the list, i.e. their parts co-occur very often.

The combinations which had not been fixed in the dictionaries before were also retrieved during the experiment. The analysis of these combinations that show both high and low values of measures of association (one or several), reveals, that bigrams which stand on the top of the list of collocations (sorted on decrease), with some degree of probability prove to be set phrases and, hence, can be included in the dictionary. The overwhelming majority of collocations that stand in the bottom part of the list prove to be free phrases.

Also it is possible to note the combinations recognized by us as collocations, but not listed in dictionaries. In case of large value of a measure for such combinations one can say to a certain degree that they belong to a class of set phrases: for example, *центр внимания* “the focus of attention”, *укромное место* “secluded corner”, *покончить жизнь* “to commit suicide”, *драконовский закон* “draconian law”, *щекотливый вопрос* “ticklish question” etc.

4. Conclusion and Further Work

The results of this work (and the data about word collocability in general based on statistical measures), first of all, can be applied to a lexicographic practice.

The statistical collocations which were extracted by measures of association, and not fixed in any dictionary, can be added to the existing dictionaries after a careful analysis. Application of corpus methods to the analysis of lexical collocability will allow to create, finally, the dictionary of a new type, namely an integrated dictionary of set phrases, or the dictionary of collocations.

It is obvious, that the automatic text analysis (for example, by means of the above described statistical tools) is only an initial stage for retrieving collocations. Then the received results must be manually processed within the framework of traditional linguistics and compared to the data from dictionaries (first of all, explanatory dictionaries and dictionaries of set phrases).

One should take into account also structural formulas which underlie collocations. Combined with statistical approaches, in our opinion, they could give quite good results. Programs which allow for stop-words and punctuation marks must also be used. Syntactic tree banks may solve the task in question. It is possible to combine statistical tools with structural (syntactic) models of phrasemes and collocations, thus, uniting two approaches.

Acknowledgements

I'm deeply grateful to my supervisor Victor Zakharov for his inspiring lectures, insightful comments, and for his support during this work. And I would like also to thank Serge Sharoff for the opportunity to work with the corpus.

References

Abramov N. (2006). *Slovar' russkikh sinonimov i skhodnykh po smyslu vyrazhenij*. Moscow.

Akhmanova O.S. (1966). Slovar' lingvisticheskikh terminov. Moscow: Sovetskaja Enciklopedija.

Bol'shoi akademicheskij slovar' russkogo jazyka. (2004-2007) (to be continued). Saint-Petersburg. Vol. 1-6. (BAS-25).

Borisova E.G. (1995a). *Kollokacii. Chto eto takoje i kak ikh izuchat'*. Moscow.

Borisova E.G. (1995b). *Slovo v texte. Slovar' kollokacij (ustojchivykh slovosochetanij) russkogo jazyka s anglo-russkim slovarem kljuchevykh slov*. Moscow.

Deribas V.M. (1983). *Ustojchivyje glagol'no-imennye slovosochetanija russkogo jazyka*. Moscow.

Fillmore C.J. (1968). The case for case. In E. Bach, R.T. Harms eds. *Universals in linguistic theory*. L. etc.: Holt, Rinehart and Winston: 1-88.

Firth J.R. (1957). Papers in Linguistics 1934-1951. Bloomington & London: Undiana University Press.

Iordanskaja L.N., Mel'chuk I.A. (2007). *Smysl i sochetajemost' v slovare*. Moscow: Jazyki Slavjanskikh Kul'tur.

Mel'chuk I.A. (1960). O terminakh "ustojchivost" i "idiomatichnost". *Voprosy jazykoznanija*. № 4: 73–80.

Sharoff S. (2002). *Chastotnyj slovar' slovar' russkogo jazyka*.
URL: http://www.artint.ru/projects/frqlist.asp

Slovar' russkogo jazyka. (1957-1961). Moscow. Vol. 1-4. (MAS).

Slovar' sovremennogo russkogo literaturnogo jazyka (1948-1965). Moscow. Vol, 1-17. (BAS-17).

Stubbs M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 1: 23-55.

Teliya V.N. (1996). *Russkaja frazeologija: semanticheskij, pragmaticheskij i lingvokul'torologicheskij aspekty*. Moscow.